Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

In the bike-sharing dataset, we explored the impact of the categorical variable 'weathersit' on the target variable 'cnt'. During exploratory data analysis (EDA), I visualized the relationship between the categorical variables and the target variable. It was observed that during ***weathersit_3*** *(Light Snow, Light Rain + Thunderstorm + Scattered Clouds, Light Rain + Scattered Clouds), there was an approximate decrease of 0.333164 units in bike hire numbers. Similarly, insights were drawn from* ***season_Spring****, while* ***season_Winter*** *displayed an opposite trend.*

Additionally, during model building, we observed a significant change in the R-squared and adjusted R-squared values when categorical features like 'yr' and 'season' were included. This suggests that the categorical features played a crucial role in explaining a greater proportion of variance in the dataset.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

When creating dummy variables (dummy encoding), it's recommended to set drop_first=True to avoid creating redundant features. Without this, the first column becomes the reference group, which can lead to correlated dummy variables.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Pair-Plot tells us that there is a LINEAR RELATION between 'temp','atemp' and 'cnt'.

The numerical variable 'registered (0.95)' has the strongest correlation with the target variable 'cnt' when considering all features. However, after data preparation, when 'registered' is dropped due to multicollinearity, the numerical variable 'atemp (0.63)' exhibits the highest correlation with 'cnt'.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Assumptions of Linear Regression:
A linear relationship exists between X and Y.
The error terms are normally distributed with a mean of zero (not X or Y).
Residual analysis of the training data shows that the residuals are normally distributed.
Therefore, our assumption for linear regression holds.
The inclusion and exclusion of independent variables in each model are based on VIF and p-values to mitigate multicollinearity.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

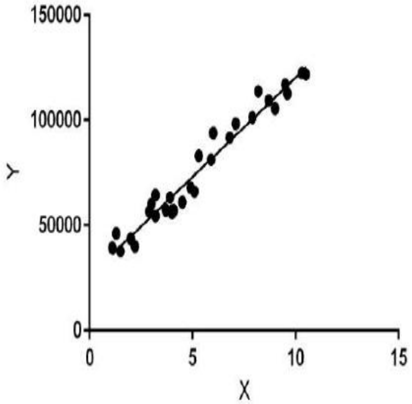
Temperature (temp) *- A coefficient value of '0.5634' indicated that a unit increase in temp variable increases the bike hire numbers by 0. 5634' units.*
Weather Situation 3 (weathersit_3) *- A coefficient value of '-0.3019' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3019 units.*
Year (yr) - *A coefficient value of '0.2310' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2310 units.*

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear regression is one of the simplest and most widely used algorithms in statistics and machine learning for predictive modeling. It is a supervised learning algorithm that is used to predict a continuous dependent variable (target) based on one or more independent variables (features). The goal is to model the relationship between the dependent and independent variables using a linear equation.
Types of Linear Regression:
Simple Linear Regression*: Involves a single independent variable and models the relationship between it and the dependent variable.*
Equation: $Y=\beta0+\beta1X+\epsilon Y$



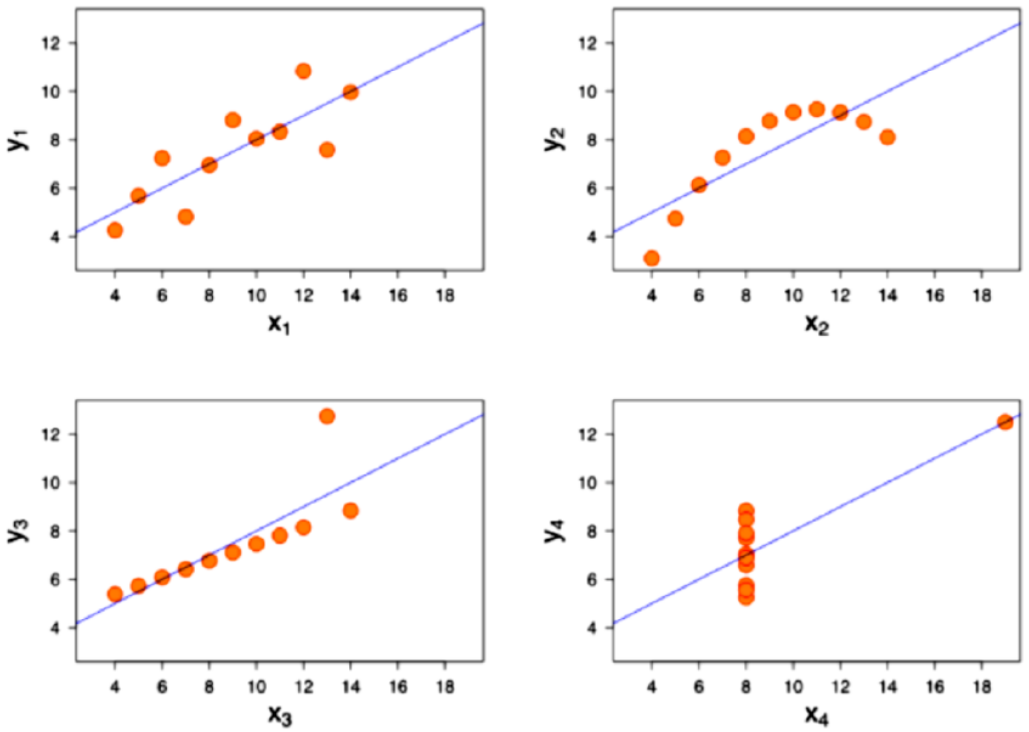
Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
***Multiple Linear Regression**: Involves two or more independent variables and models the relationship between them and the dependent variable.*
Equation: $Y=\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta nXn+\epsilon Y$

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>
Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics but very different distributions and relationships between the variables.

Anscombe's Quartet consists of four datasets: X1, Y1, X2, Y2, X3, Y3, X4, Y4. Each dataset has 11 data points, and the datasets share the following characteristics:

- *The mean of X and Y is the same in all four datasets.*
- *The variance of X and Y is the same in all four datasets.*
- *The correlation between X and Y is the same in all four datasets.*
- *The regression line (least-squares fit) is the same in all four datasets.*



Summary Statistics for Anscombe's Quartet:

Here are the summary statistics for all four datasets (mean, variance, correlation, and regression line):

Dataset	*Mean of X*	*Mean of Y*	*Variance of X*	*Variance of Y*	*Correlation (r)*	*Linear Regression Equation (Y = a + bX)*
1	*9*	*7.5*	*11*	*3.25*	*0.816*	*Y = 3 + 0.5X*
2	*9*	*7.5*	*11*	*3.25*	*0.816*	*Y = 3 + 0.5X*
3	*9*	*7.5*	*11*	*3.25*	*0.816*	*Y = 3 + 0.5X*
4	*9*	*7.5*	*11*	*3.25*	*0.816*	*Y = 3 + 0.5X*

Anscombe's Quartet serves as a powerful reminder that "data tells a story," and the process of analyzing data should always include visualization. Even if your data summary statistics seem similar, different data structures can lead to different interpretations and conclusions.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

***Pearson's R**, also known as the **Pearson correlation coefficient** or **Pearson's product-moment correlation**, is a statistical measure used to evaluate the strength and direction of the linear relationship between two continuous variables. It is one of the most widely used methods to quantify the degree of correlation (or association) between two variables.*

Formula:

The formula for Pearson's R is:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Where:

- *R is the Pearson correlation coefficient.*
- *Xi and Yi are the individual data points of the two variables XXX and YYY.*
- *X¯ and Y¯ are the mean values of the variables XXX and YYY, respectively.*
- *The summation (∑) runs over all data points.*

Suppose you have two variables, X = [1, 2, 3, 4, 5] and Y = [2, 4, 6, 8, 10].

- *Mean of X = (1+2+3+4+5)/5 = 3*
- *Mean of Y = (2+4+6+8+10)/5 = 6*
- *Using the Pearson formula, we would calculate the correlation coefficient, which in this case would be r = 1, indicating a perfect positive linear relationship.*

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

***Scaling** is the process of transforming features of your data so that they have specific properties, such as a particular range or distribution, in order to make them more suitable for machine learning algorithms. Many algorithms work better or converge faster when the features are on a similar scale. Scaling helps to normalize the data by adjusting the range and distribution of variables.*

In simple terms, scaling ensures that all features contribute equally to the analysis or model, preventing features with larger numerical ranges from dominating the learning process.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

- *It brings all of the data in the range of 0 and 1.*
- *sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.*

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- *Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).*

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- *sklearn.preprocessing.scale helps to implement standardization in python.*
- *One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers*

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

The Variance Inflation Factor (VIF) is a measure of colinearity among predictor variables within a multiple regression. It is calculated by taking the the ratio of the variance of all a given model's betas divide by the variane of a single beta if it were fit alone.
If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

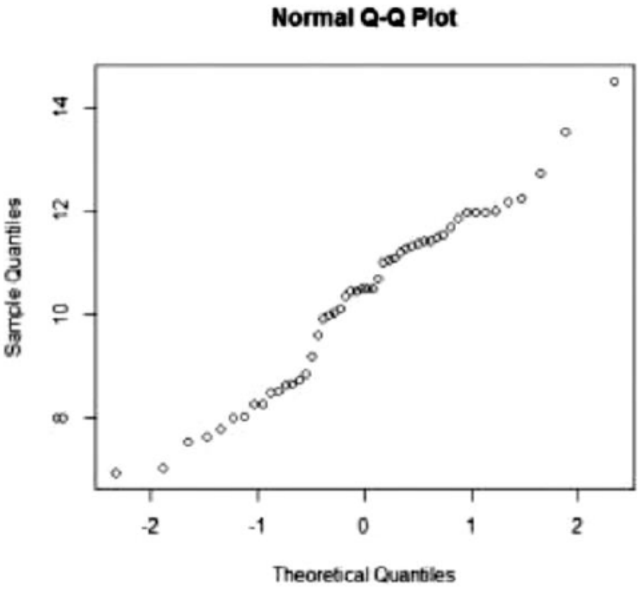
Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>
Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.
<u>Few advantages:</u>
a) It can be used with sample sizes also
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios: If two data sets.
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight i.e.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles